Advanced conceptual network usage in library database queries

Hoppenbrouwers, J.J.A.C.

Publication date:
1998

Advanced Conceptual Network Usage in Library Database Queries

Jeroen Hoppenbrouwers
Infolab, Tilburg University

September 1998

Abstract

This paper describes the generic principles of Decomate-II's Concept Browser. It discusses the three main problems in using a thesaurus or conceptual network to help users formulate database queries: the network maintenance problem, the network navigation problem, and the network-to-database mapping problem. Possible solutions to all these problems are proposed, based on previous experiences with concept network systems. Care is taken to keep the resulting system suitable for production use in an existing library environment based on Boolean keyword retrieval from a large collection with uncontrolled vocabulary.

Keywords: Semantic network, thesaurus, conceptual network, knowledge navigation, lexicon, conceptual modeling, topic browsing, document retrieval, Decomate-II.

URL: `http://infolab.kub.nl/prj/decomate`

1 Introduction

Much work in the area of indexing and retrieval concentrates on constructing effective and efficient algorithms to find a set of 'interesting' documents in a large collection, given a user query of some sort. So-called *full text query engines* are used to mechanically select documents out of a collection, either by Boolean keyword matching or according to statistical computations (mainly clustering techniques). Boolean engines usually return all and only documents that exactly match the query. The statistic retrieval engines use various combinations of text metrics in order to predict the document's value in terms of the user query (Salton and McGill, 1983; Salton, 1991). This is called *relevance ranking*. Usually only a limited amount of documents (e.g., the top 100 best matches) is returned to the user.

1.1 The Keyword Barrier

Because most relevance ranking algorithms are based on *textual* data, i.e., actual word forms (strings) without any sense interpretation, they tend to perform generally mediocre (Woods, 1997; Shaw et al., 1997; Blair and Maron, 1985). The same problem appears in Boolean engines. Blair and Maron list many problems with keyword-based searches, and most of their findings fall into two categories:

- Phraseology: synonyms, slang, and jargon terms obscure the meaning of the text, making it very difficult to locate by keyword approaches.
- Granularity: as database size increases, increasingly fine-grained searches are necessary. Retrieval techniques that only consider the presence or absence of words cannot distinguish different relationships between the same words, and retrieve far more documents than the user considers relevant (Carbonell and Thomason, 1986).

The crucial problem of current information retrieval technology is that systems relying solely on the presence or absence of a word are inherently limited in their ability to distinguish relevant and irrelevant texts. Word-based systems that cannot deal with synonymy, polysemy, metaphor, and the other complications of natural (uncontrolled) language must have some upper bound on retrieval performance, the *keyword barrier* (Maulding, 1991).

One way of breaking the keyword barrier would be to add semantic knowledge about the document content to the database, so that the above (mostly syntactic and lexical) problems are avoided. Unfortunately, this is not feasible in the context of the Decomate-II project, since the databases and query engines are given. Our only chance to break the barrier is to enrich the query that is sent to the database backends, using a 'knowledgeable' intermediary that *does* possess semantic knowledge, and treat the returned results with a comparable device for additional filtering, to provide relevance feedback, and to add relevance ranking.

The usage of so-called *search intermediaries*, either human (librarians or documentalists) or mechanical, has been proposed both to shield the user from the technical aspects of the query engine and to provide extra background information about the document's domain (Wiesman, 1998). The objective of these systems is to be helpfully responsive to a spontaneous description of the information required, minimizing the need for an information seeker to engage in repeated query reformulation in order to discover the exact terminology that will retrieve the information required (Hoppenbrouwers, 1998; Woods, 1997, p.12).

This paper focusses on various difficulties in applying such a search intermediary to an *existing* library system, based on Boolean keyword retrieval from an uncontrolled vocabulary.

1.2 The Concept Browser in Decomate-II

The Decomate-II Library System, currently under development at Tilburg University in cooperation with several European partners, aims at a Web-based single point user interface to a multitude of (possibly distributed) databases. A single user query, usually a set of keywords, is mapped to all connected databases, each with its own query language, data schema, and content. The individual query results are merged together by the system, post-processed to eliminate noise such as double documents, and presented to the user in a suitable format.

Decomate-II is a follow-up to Decomate-I (Dijkstra, 1998b; Dijkstra, 1998a), which aimed mainly at desktop delivery of (copyrighted) material through electronic channels. Just as Decomate-II it provided a Web-based user interface, but it lacked the integrated single entry point, and the distributed query and result merging capabilities.

An additional feature that is part of Decomate-II is the *Concept Browser*, which is built on top of the aforementioned database engines. See Figure 1 for an illustration of the Decomate-II architecture. The standard query system is directly from a user's browser to the Broker (through a Web server, which is not indicated in the figure). A user can also select to use the Concept Browser, which will present him or her with a conceptual map of the relevant domain (in Decomate-II, Economics). The user can then navigate the domain, while the Concept Browser collects information about chosen paths and nodes. Eventually the Concept Browser generates a normal query which is passed to the Request Broker.

Broker results can be directly passed back to the user, or be post-processed through the *Concept Mapper*. This module, which will not be discussed further in this paper, will be able to augment the query results with semantic information and facilitate document grouping, relevance ranking, and relevance feedback (Fitzpatrick and Dent, 1997; Salton and Buckley, 1990). It is our intention to make the Concept Mapper also available for direct queries to the Broker, i.e., without first going through the Concept Browser. However, the amount of extra information available after a Concept Browser session might significantly increase the effectiveness of the Concept Mapper, especially in the area of relevance ranking. Note that the database engines themselves do not provide relevance ranking, if any ranking at all, and that the ranking results of separate databases are very difficult to compare and merge.

We currently plan to implement the Concept Browser and Concept Mapper as distributed client-server systems, with some code executing on the Decomate-II server and some (Java) code in the user's Web browser. Most likely the thesaurus/conceptual network will remain on the server, while the GUI is downloaded to the client and connects back to the server, to retrieve relevant concepts on demand and to execute the actual query. This architecture has successfully been used in comparable systems, such as IBMs *Lexical Navigation* (Cooper and Byrd, 1997), and should be flexible

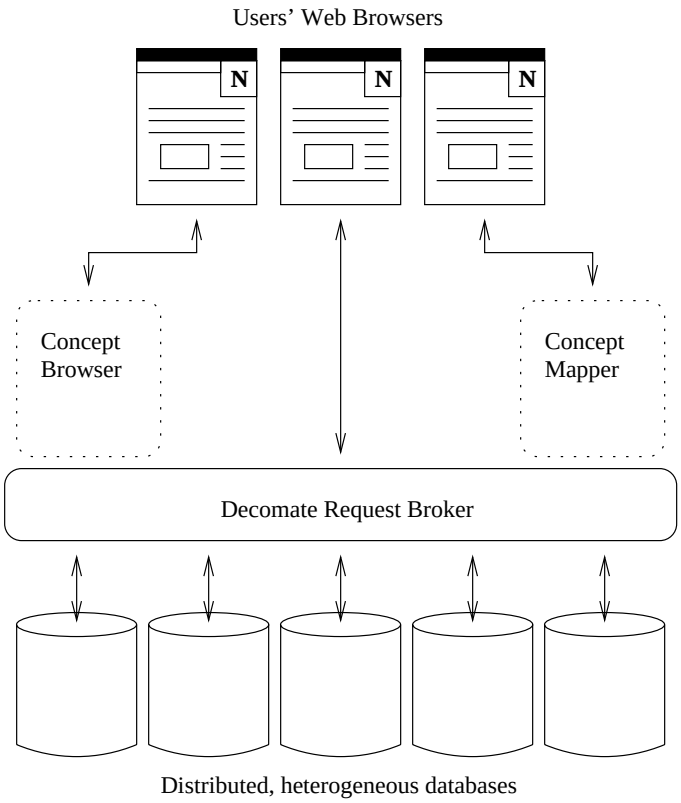

Figure 1: The Decomate-II Architecture

and scaleable enough for large-scale production work.

Several implementation and operational restrictions complicate the addition of concept browsing and mapping modules to Decomate-II (Hoppenbrouwers, 1998):

- Wide variation in underlying databases (query language, available information, level of detail, result ranking...).
- Need to integrate with existing system components, even when they are not 100% suitable. This holds especially for the various databases, some of which are outside library control (e.g., CD-ROM databases provided by third parties).
- Availability of Boolean keyword queries only; no statistical (e.g., vector, cluster, or frequency) data on the databases is or will become available.
- Restrictions on available system capacity, in particular processing and network bandwidth, with firm limits on allowable response time (tak-

ing into account that database queries may already take up lots of time).

- Limited conceptual network creation and maintenance capacity by librarians and documentalists.
- Possibly distributed conceptual networks which should be connected or duplicated.

Especially the given database structure sets this project apart from earlier attempts; we do not have the possibility to create a well-tuned, dedicated database engine with state-of-the-art index and search techniques. Furthermore we emphasize the usage of *purpose-made conceptual networks* for the Browser Interface, while other projects (Cooper and Byrd, 1997) often use machine-generated lexical networks. We believe that it is more likely to get satisfying results by keeping the network in the hands of qualified professionals. In a sense, we do not aim to produce a virtual library, but a virtual librarian. Lastly, there should be provisions to partition the conceptual networks to facilitate load balancing, knowledge balancing (specialized networks), and maintenance balancing. Linking conceptual networks in a federated manner is not trivial.

1.3 Conceptual Queries and Knowledge Navigation

Parsaye *et al.* (1989, Ch.6) distinguish five kinds of knowledge that users need to have in order to successfully compose conceptual queries for information retrieval systems: procedural system knowledge, strategic search request formulation knowledge, indexing policy knowledge, search strategy knowledge, and domain knowledge. Of these five, the first four types of knowledge can be leveraged by suitable user training and system manuals. Domain knowledge, however, is typically not suitable to be presented in a manual or through simple trainin, and when extensive education is not an option, some form of extra help must be supplied by the system. Additionally, extra domain knowledge is required to formulate good queries, especially with complex information needs (Hoppenbrouwers, 1998; Bodner and Song, 1996; Howard, 1992).

Conceptual queries involve *knowledge navigation*. There are two basic types of knowledge navigation: *searching* and *browsing* (Wiesman, 1998, p.7). Since searching implies that the user already knows what to look for, this is no solution to the lack of domain knowledge. In such a situation, complete *topic-level concept browsing* is required (Hoppenbrouwers, 1998; Papazoglou, 1997). Interactive term suggestion, where the system suggests terms for the user to choose, can also significantly enhance retrieval effectiveness (Schatz et al., 1996; Papazoglou et al., 1998).

Basic idea behind such a system is to provide the user with an 'information space' (IS) through which (s)he can freely browse, and which gives a fair indication of the *distribution of concept terminology* over the domain.

Although the IS should have a reasonable overlap with the actual database, it only serves as a guide: actual queries still are served by the underlying original databases. This means that the concept space browser is allowed to be slightly out of date without significantly affecting the final query result.

In an extreme case, the conceptual space could be very extensive with a very limited document database, a situation comparable with an experienced librarian or documentalist who starts up a new, therefore small, library. The other extreme, a small conceptual space with a large collection of documents, represents a well-stocked library with few support people who know what the documents actually are about.

1.4 Thesauri and Conceptual Networks

There are several descriptions and definitions of conceptual networks, sometimes called *ontologies*, *thesauri*, *lexicons*, or *semantic fields* (Hoppenbrouwers, 1997). Traditionally the difference between these and other related terms is the following.

The vocabulary provides the official list of correct forms of words, presents only syntactical features, and gives idiomatic patterns of usage if necessary (Weigand, 1990, p.77).

A thesaurus provides the official survey of correct terminology for concepts, presents only basic semantic features, structures the terms in a semantic net, and adds to the vocabulary special patterns of usage appropriate to the special concepts (Weigand, 1990, p.77). Alternatively, a thesaurus is the vocabulary of a controlled indexing language, formally organized so that *a priori* relationships between concepts are made explicit (Aitchison and Gilchrist, 1987).

A dictionary presents the definitions of terms from the thesaurus, gives humans the understanding of specialized words, helps shape the growth of the thesaurus, and helps authorities in deciding on the admission of new terms (Weigand, 1990, p.77).

An ontology is 'a systematic account of Existence,' a description of the minimal set of concepts that a language needs to express all its other concepts (Kaminsky, 1969). Pragmatically, an ontology defines the vocabulary with which queries and assertions are exchanged among agents (Gruber, 1993).

A lexicon combines the vocabulary and the thesaurus and integrates them in a machine-readable format so that it can be managed and queried by computation engines (Hoppenbrouwers, 1997).

A semantic field or *domain* is a more or less clearly outlined subsection of the real world. It can often be related to a group of people who live and work in the same environment, a *semantic community* (Robinson

and Bannon, 1991; Ulijn and Strother, 1995). Some authors claim that within any semantic field, there must be no ambiguous terminology (Wiederhold, 1995), thereby putting a very hard constraint on the semantic field up to the point that only one single person can live in such a field at the same time.

A conceptual network is a collection of semantic nodes with links between them, in such a way that many relationships are captured. Covering a semantic field, it is usually much more extensive than a typical thesaurus, e.g., containing semantic roles and part-of relationships (Miller et al., 1993). However, newer thesauri contain more and more information and can be assumed to be conceptual networks as well (Miller, 1997).

Note that most 'strict' ontologies or thesauri that propose one single hierarchical organisation are bound to fail (Sowa, 1983, p.15); no linguistic or psychological evidence has uncovered a truly universal set of primitives, and often it is difficult if not impossible to assign concepts to only one category (Woods, 1997). Likewise, because the vocabulary of each living language grows with approximately 6000 lemmas per year (Ulijn and Strother, 1995, p.101), especially in the technical-scientific register, it will be very hard to claim that *any* thesaurus is ever complete. Regular updates must be applied to every thesaurus to keep it in synchronization with the evolving semantic field (Aitchison and Gilchrist, 1987).

The dynamic nature of thesauri and conceptual networks means that mostly static, hierarchically organized classifications such as the UDC tree[1] or the classification of the Journal of Economic Literature (JEL)[2] are not sufficient to serve as a complete conceptual network. Besides, they do not aim at covering the terminology of the semantic field—they want to identify specific subfields (subjects) within the larger fields. Of course their *subject headings* can be used as a starting point for thesaurus construction, and they can be included as generic 'see also' pointers in a conceptual network.

Conceptual networks such as WordNet (Miller et al., 1993) contain enough terminology and relationship information to be usable, however, they usually are too static as well and cover a broad range of common semantic fields while being sparse on detailed, specialist fields—which are far better suited to assist users in knowledge navigation (Bodner and Song, 1996; Howard, 1992). It is especially important to have the conceptual network organized in terms of, indeed, *concepts*, instead of plain terms. WordNet uses the *synset* primitive to group highly synonymous terms together, and the EuroWordNet Project extends the synonymity relation to include multiple languages (Vossen, 1997; Vossen et al., 1997). Other work on Lexicons, aimed specifically at conceptual modeling (Hoppenbrouwers,

[1] `http://main.bib.uia.ac.be/MAN/UDC/udce.html`
[2] `http://www.econlit.org/elclasbk.htm`

1997), also suggests ways of organizing terminology to properly present a conceptual space to users.

Acquiring a suitable conceptual network therefore is not just a matter of copying existing thesauri or term lists. Considerable effort should be put into the initial creation of a conceptual network for knowledge navigation purposes. This is not to say that existing information cannot be reused, but it usually requires extensive post-processing.

Besides the acquisition problem, there is also a maintenance problem, a navigation problem, and a mapping problem. The next sections will consider each problem in more detail, suggesting possible solutions as appropriate.

2 Maintenance Problem

Any semantic network which models a piece of the world needs regular updating in order to stay synchronized with the world. The idea that a network could be constructed once and remain stable for an extended period of time should be abandoned:

> The danger is that if the thesaurus is permitted to become monolithic and resistant to change, it can actually hinder both indexing and retrieval.
>
> (Milstead, 1992)

There are two separate groups of people who naturally should get involved in conceptual network construction and maintenance: library professionals (documentalists and librarians) and research professionals (scientists) who regularly use the library.

2.1 The Role of the Documentalist

In case of a library system which specializes in one particular scientific field, such as Economics, the network should be maintained by experienced documentalists who are comfortable with this field. These people can quickly recognize the particular spots in the network where potential new concepts should be placed, and can update and use the network as part of their regular work. In this way they develop a 'map' of their field, which can be very useful for other purposes besides knowledge navigation support. We assume here that documentalists are explicitly keeping up with the scientific field; not that they just catalog new publications. Only then, proper maintenance of the conceptual network is guaranteed.

Of course, it should be assured that network maintenance is a technically simple operation that does not need much training or time, and that the immediate day-to-day advantage for the documentalists is sufficiently

clear. Only then will these people be inclined to spend effort on network maintenance. This is an important part of the whole project, because without network maintenance, the relevancy of the concept browser will diminish over time.

Automated tools should be available to help documentalists with network maintenance as much as possible, e.g., by proposing specific places in the existing network for new concepts. The documentalist then might only need to confirm the system's proposal. In addition, tools for network manipulation (moving, copying, deleting, printing) should be readily available. A good browser which exceeds the simple record-oriented concept view (a single concept with all associated direct links only) is also required, in order to facilitate situational awareness.

2.2 The Role of the Scientist

It is a safe assumption that the same scientists who use the library have at least a partial task in providing its contents as well. Not only do they influence the collection (although usually only in an indirect way), they also publish documents about the same scientific field. This implies an intimite knowledge of at least a part of the field.

Leaving all the network maintenance to documentalists and librarians denies the obvious knowledge that some library users already have. Especially during the inital network construction, library people should consult researchers in order to build up a coherent and reasonably complete network. Note that it is not required that all scientists fully agree on the network; it is no standard classification. The purpose of the network is to assist less experienced people, and when several 'research schools' exist, that fact should be noted, not voted away.

3 Navigation Problem

The networks (thesauri) typically used in information retrieval systems are simple more generic/more specific hierarchical trees, sometimes extended with cross links such as 'used for' and 'related terms.' Synonyms may be present, but the main purpose of classification hierarchies usually is to avoid any synonymous references.

Obvious links such as between 'software engineering' and 'software protection' are often missing, even when both terms are present, because they do not typically classify under the same head word. As an example, in the thesaurus used by Excerpta Informatica at Tilburg University, 'software protection' was classified directly under 'software,' but 'software engineering' was generalized by 'software technology,' which was not at all mentioned under 'software' but was located under 'computer technology.' In other words, there was no navigational path between the terms, even though they both started with the same word. Because the Excerpta

thesaurus could also be queried as if all terms together were a full text database, the 'software' keyword in 'software engineering' was found, but the result of this query was an alphabetically ordered list of 61 items which did not all contain the 'software' search term.

It is a well-known problem of any hierarchically organized system (even when cross links are present) that concepts often do not naturally classify under one single category (Woods, 1997). Concepts should be placed in multiple locations in the network, participating in several tree structures if required. However, when users navigate a particular tree, it should be made clear to them when they are about to leave their original tree.

When fully associative lexicons are navigated through a suitable interface, such as the various WordNet interfaces (Miller et al., 1993), link types such as more generic/more specific are usually made explicit. Other variants exist that use one single link type for all links, but give these links different 'weights' to indicate the 'strength' of the link between two concepts.[3]

The whole issue of hierarchical navigation boils down to the difference between structural (analytical) browsing and associative browsing. Despite the obvious maintenance and implementation advantages of strict hierarchical concept trees with explicit cross links outside the tree structure, an associative structure is better suited to model a typical semantic field. Psycholinguistically inspired lexicons such as WordNet therefore offer more link types, like synonyms, antonyms, meronymy, and others (Miller et al., 1993). Presenting these links to the user in an easy format is not trivial; most likely, some form of two-dimensional graphical browser is needed. Considerable work has been done in this area, see Aitchison (1987) and Lancaster (1986).

Many helpful ideas have come from the Lexical Navigation Project (Cooper and Byrd, 1997), where a browser client was developed to facilitate user browsing of a lexical network. Although there are differences between an lexical and a conceptual network, these differences are not necessarily visible to the user. The local, non-persistent network management that can be done, such as moving nodes on the screen or deleting nodes to create a specialized view or subnet, seems very helpful to browing and query creating activities.

4 Mapping Problem

Some current thesaurus-assisted information retrieval systems use the thesaurus to let the user navigate to a certain term, and subsequently use only this term to query the (full-text) database. Although this method certainly helps to suggest particular search terms to the user, it does not at all use the semantic network formed by the links between words to improve the

[3] `http://www.plumbdesign.com/projects/thesaurus.html`

query, i.e., to increase the system's recall and precision.

The semantic network actually serves two distinct purposes. First, it helps the user to become familiar with a certain semantic field which (s)he might not fully grasp yet. In this way, the network might help the user to come up with relevant terms, which will yield better results than forcing the user to key in keywords from the top of his/her head. Second, the network offers the system the option of adjusting the user query by not using only the term the user indicated, but also using the terms around the user term and maybe even other concepts. Using more of the network effectively links the thesaurus with the library database instead of just using the thesaurus for term suggestions, then throwing all thresaurus link information away and start over with a clean query (cut and paste approach).

Candidate thesaurus terms to be called in by the query generator are synonyms (although care must be taken not to expand the query beyond what the user intends), direct hypernyms, some levels of hyponyms, and maybe some levels of meronymic relationships (Maulding, 1991). Note that these are *candidate* terms, not necessarily actual terms used for query expansion. Especially synonyms usually pose problems, as the following quote from Woods explains:

> A common approach to the paraphrase problem is to use tables of synonyms to automatically expand queries by adding terms that are recorded as 'synonymous.' However, there are few real synonyms in English, so the common practice is to include related words as if they were synonyms. However, treating terms this way when they are not really synonyms introduces a level of granularity that trades off precision for recall. There is no a priori correct level for this tradeoff—different information needs require different levels of generality—so this technique often degrades retrieval rather than improving it.
>
> As an alternative to synonym classes, we use taxonomic subsumption algorithms that exploit generality (subsumption) rather than synonymy to connect terms in queries with passages that contain more specific terms as well as the requested terms. These algorithms do not automatically explore more general terms, so the level of generality is controlled by your choice of query terms. For example, if you ask for 'motor vehicles' you would get trucks, buses, cars, etc., but if you ask for 'automobiles' you would get cars and taxicabs, but not trucks and buses.
>
> Using knowledge bases of general semantic facts, structured conceptual taxonomies (a type of semantic network) can be constructed from words and phrases. These words and phrases can be extracted automatically from text and parsed into conceptual structures. The taxonomy can be organized by the most-specific-subsumer (MSS) relationship, where each concept is linked to the

> most specific concepts that subsume it—i.e., that are more general than it is. Terms in a query are individually matched with corresponding concepts in the taxonomy together with their subconcepts.
>
> For example, given the general semantic facts that 'washing' is a kind of 'cleaning' and 'car' is a kind of 'automobile,' an algorithmic classification system can automatically classify 'car washing' as a kind of 'automobile cleaning.' A query for 'automobile cleaning' or 'automobile washing' will immediately retrieve hits for 'car washing.'
>
> (Woods, 1997)

More than the above terms are not available when the user came to a term without any navigation, i.e., by hitting a spot on the 'map' and not moving from there. When the user has navigated the map, however, much more information is available to construct a query.

For example, a common way to end up at a term is to follow one of the available tree structures down from the top. There certainly is a noticeable difference in arriving at 'software engineering' via 'technology' and 'engineering' than arriving via 'computer programming,' 'modular software,' 'component reuse' or even through an associative sidelink as in 'organization,' 'business process redesign,' 'business process re-engineering,' 'engineering.' Traditional thesauri go to great efforts to prohibit multiple paths to concepts, but a well-designed network might provice invaluable extra clues to determine the nuances and context the final term should be interpreted in (Aslandogan et al., 1997).

The ability for the user to mark terms while passing them by, so that the resulting query will take the marked words into account as well, can also enhance the standard 'query for this term' feature.

5 Preliminary Design of the Concept Browser

What eventually must be envisioned is an interface where a user can browse the complete semantic field in an intuitive way, while the system follows his/her traces in order to compile a good query for the actual library database. This would involve using both implicit and explicit user signals in order to collect the required knowledge. For example, taking certain 'turns' in the network, circling around a concept, and other implicit navigational actions give clues to the underlying 'knowledge snooper' what the user might be interested in. Explicit cues, e.g., by marking certain concepts as being of special interest, then provide stronger input to the snooper and might be used to actively propose certain new pathways.

The browser interface in this way assumes almost the role of an *agent*, a semi-intelligent piece of software that plays the role of a *virtual librarian*,

guiding the user through the massive amount of knowledge available in the conceptual network (and eventually in the underlying library).

Other areas of interest are history extrapolation through machine learning (Chen, 1995), specifically linked example documents, interactive term and field suggestions (based on previous queries from other people (Fitzpatrick and Dent, 1997), but reviewed by a documentalist), and direct user-machine dialog to create 'pre-query relevance feedback' (Cooper and Byrd, 1997).

Placing the user somewhere in the conceptual network to start browsing is not trivial. A common way of finding a suitable starting place in keyword systems is by having the user enter an initial query. However, as argued above, there usually are more places in the same network that warrant attention. Moreover, it is not a good idea to restrict database queries to only the controlled terminology in the network. After all, most entries in the database consist of free natural language, especially if the database contains abstracts and/or full text. When queries are only allowed on keywords from the restricted network, they will in many cases miss important documents. On top, the initial user suggestion itself might contain important terms that are not obvious in the domain. This implies that a good conceptual network should contain as many terms as possible, and be not restricted to 'preferred' terminology (although some concepts certainly may be flagged as 'preferred' for explicit keyword assignments to database entries). Even words not traditionally associated with a given semantic field should be included to enable intuitive associations to be made.

For example, a user looking for generic 'changes in the deficit' would expect a document to turn up which contains 'Last year's reductions in tax rates are part of the reason for the deficits, as are the administration's plans for a sustained military buildup.' While the query and the passage convey similar ideas, the wording in each is different, a typical case of the paraphrase problem. 'Change' is a more generic variant of both 'reduction' and 'increase,' but neither of the three would be expected in any pure economics thesaurus. Without these terms, either far too many documents would show up, or none at all (Woods, 1997).

This means that the network presentation in the user interface must be dynamic, and must allow navigation in several places at the same time, 'connecting' parts of the network that are not connected by default (Cooper and Byrd, 1997). Additionally, there are technologies which can map a user-provided term into the thesaurus even if the thesaurus does not contain that term explicitly (Woods, 1997). This can help to position the user on the semantic map in several promising starting places, plus to modify the map itself when there is need to do so.

A great deal of effort must be put into an attractive user environment, which makes users sufficiently comfortable to spend some time on query refinement. This is so important because numerous research projects (Lesk,

1998; Jansen et al., 1998; Fitzpatrick and Dent, 1997) indicate that typical query engine users heavily prefer extremely short queries (1.5 word on average).[4] People are lazy and rarely try feedback on raw queries (*More Like This*) or read the instructions, even if this would give them huge improvements on their query's success. They do not use many Boolean operators either and rarely understand them, and have a tendency to use a limited amount of only 'standard' terms in their queries.

As Jansen et al. conclude, the low use of advanced search techniques would seem to support the continued research into new types of user interfaces, intelligent user interfaces, or the use of software agents to aid users in a much simplified and transparent manner. This is exactly what Decomate-II's Concept Browser attempts. The Concept Mapper will attempt to improve usage of relevance feedback in much the same way.

6 Conclusions and Future Work

This paper is not intended to provide complete solutions to the observed problems. We hope to have indicated some promising ways to enhance Boolean keyword queries compiled by non-professionals *before* the queries are presented to the database engine(s). Several ideas from earlier research and previous implementations have been included in the preliminary design of a Concept Browser.

What must be done now, in context of the Decomate-II Project, is to draft a complete design for such a system, including the selected solutions for the acquisition, maintenance, navigation, and mapping problems. This design must fit in with existing Decomate modules, especially the Multi-Protocol Server (de Cock, 1998) and the User Interface Generator.

Within the time frame of the project, we plan to build a fully operational prototype of the Concept Browser and Concept Mapper, and aim to finish production versions near the end of the project in July 2000. Preliminary user evaluations and usage logs will be used for intermediate interface correction suggestions.

References

Aitchison, J. and Gilchrist, A. (1987). *Thesaurus Construction*. Aslib, London. 2nd edition.

Aslandogan, Y., Thier, C., Yu, C. T., Zou, J., and Rishe, N. (1997). Using Semantic Contents and WordNet in Image Retrieval. In Belkin, N. J., Narasimhalu, A. S., and Willett, P., editors, *Proceedings of the 20th ACM SIGIR Conference*, pages 286–295.

[4] These numbers are based on typical Web search engine queries. But since the Decomate-II system will be Web-based, it seems prudent to assume that people will treat it as a standard Web query engine.

Blair, D. and Maron, M. (1985). An evaluation of retrieval effectiveness for a full-text document retrieval system. *Communications of the ACM*, 28(3):4–22.

Bodner, R. and Song, F. (1996). Knowledge-based approaches to query expansion in information retrieval. In *Lecture Notes in Computer Science*, volume 1081, pages 146–158.

Carbonell, J. and Thomason, R. (1986). Parsing in biomedical indexing and retrieval. In *AAMSI-86*.

Chen, H. (1995). Machine Learning for Information Retrieval: Neural networks, Symbolic Learning and Genetic Algorithms. *Journal of the American Society for Information Science*, pages 194–216.

Cooper, J. W. and Byrd, R. J. (1997). Lexical Navigation: Visually Prompted Query Expansion and Refinement. In Allen, R. B. and Rasmussen, E., editors, *Proceedings of the 2nd ACM International Conference on Digital Libraries*.
`http://www.ibm.research.com/people/j/jwcnmr/`.

de Cock, R. (1998). Frankenstein Returns. Personal communication, Decomate-II Project.

Dijkstra, J. (1998a). Journals in Transition: From Paper to Electronic Access: The Decomate Project. *Serials Librarian*, 33(3/4):243–270.

Dijkstra, J. (1998b). Objectives, Results, and Conclusions of the European DECOMATE project. In Barth, A., Breu, M., Endes, A., and de Kemp, A., editors, *Lecture Notes on Computer Science: Digital Libraries in Computer Science, the Medoc Approach*, pages 231–238. Springer Verlag, Berlin, Heidelberg.

Fitzpatrick, L. and Dent, M. (1997). Automatic Feedback using Past Queries: Social Searching? In Belkin, N. J., Narasimhalu, A. D., and Willett, P., editors, *Proceedings of the 20th ACM SIGIR Conference*, pages 306–313.

Gruber, T. (1993). A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199–220.

Hoppenbrouwers, J. (1997). *Conceptual Modeling and the Lexicon*. PhD thesis, Tilburg University.
`http://infolab.kub.nl/people/hoppie`.

Hoppenbrouwers, J. (1998). Browsing Information Spaces. In Prinsen, J., editor, *International Summer School on the Digital Library 1998*, Tilburg, The Netherlands. Ticer B.V.
`http://infolab.kub.nl/people/hoppie`.

Howard, H. (1992). Measures that discriminate among online searchers with different training and experience. *Online Review*, 6:315–327.

Jansen, B. J., Spink, A., Bateman, J., and Saracevic, T. (1998). Real life information retrieval: A study of user queries on the web. *SIGIR Forum*, 32(1):5–17.

Kaminsky, J. (1969). *Language and Ontology*. Southern Illinois University Press.

Lancaster, F. (1986). *Vocabulary Control for Information Retrieval*. Information Resources Press, Arlington VA.

Lesk, M. (1998). "Real World" Searching Panel at SIGIR '97. *SIGIR Forum*, 32(1):1–4.

Maulding, M. L. (1991). *Conceptual Information Retrieval: a case study in adaptive partial parsing*, volume 152 of *Kluwer international series in engineering and computer science*. Kluwer Academic Publishers.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1993). Introduction to wordnet: An on-line lexical database. Technical report, Princeton University.

Miller, U. (1997). Thesaurus Construction: Problems and their Roots. *Information Processing and Management*, 33(4):481–493.

Milstead, J. (1992). Methodologies for subject analysis in bibliographic databases. *Information Processing and Management*, 28:407–431.

Papazoglou, M. (1997). Knowledge Navigation and Information Agents: Problems and Issues.

Papazoglou, M., Weigand, H., and Milliner, S. (1998). TopiCA: A Semantic Framework for Landscaping the Information Space in Federated Digital Libraries.

Parsaye, K., Chignell, M., Khoshafian, S., and Wong, H. (1989). *Intelligent Databases: Object-Oriented, Deductive Hypermedia Technologies*. John Wiley and Sons, New York, NY.

Robinson, M. and Bannon, L. (1991). Questioning representations. In Bannon, L., Robinson, M., and Schmidt, K., editors, *Proceedings of the Second European Conference on Computer-Supported Cooperative Work*, page 219.

Salton, G. (1991). Developments in Automatic Text Retrieval. *Science*, 253.

Salton, G. and Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297.

Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, NY.

Schatz, R. et al. (1996). Interactive Term Suggestion for Users of Digital Libraries. In *1st ACM International Conference on Digital Libraries*, pages 126–133. Bethesda. MD.

Shaw, W., Burgin, R., and Howell, P. (1997). Performance standards and evaluations in IR test collections: vector space and other retrieval models. *Information Processing and Management*, 33(1):15–36.

Sowa, J. F. (1983). *Conceptual Structures, information processing in mind and machine*. Addison-Wesley, Reading, Massachusetts.

Ulijn, J. M. and Strother, J. B. (1995). *Communicating in Business and Technology: From Psycholinguistic Theory To International Practice*. Peter Lang GmbH, Frankfurt.

Vossen, P. (1997). EuroWordNet: a multilingual database for information retrieval. In *Proceedings of the DELOS workshop on Cross-language Information Retrieval, March 5-7, 1997, Zürich*.

Vossen, P., Diez-Orzas, P., and Peters, W. (1997). The Multilingual Design of the EuroWordNet Database. In *Proceedings of the IJCAI-97 workshop Multilingual Ontologies for NLP Applications, August 23, 1997, Nagoya*.

Weigand, H. (1990). *Linguistically Motivated Principles of Knowledge Base Systems*. Foris Publications, Dordrecht, Holland.

Wiederhold, G. (1995). Value-added mediation in large-scale information systems. In *Database Applications' Semantics, IFIP TC-2 DS-6*.

Wiesman, F. (1998). *Information Retrieval by Graphically Browsing Meta-Information*. PhD thesis, Maastricht University, The Netherlands.

Woods, W. A. (1997). Conceptual Indexing: A Better Way to Organize Knowledge. Technical report, Sun Microsystems Laboratories. `http://www.sunlabs.com/technical-reports/1997/abstract-61.html`.